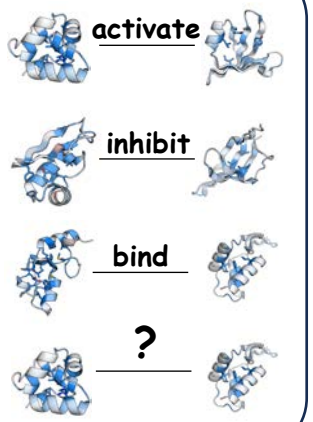
Sector 1: Protein data to natural language in ProCoT format

Protein data



Natural language in ProCoT format

Question:

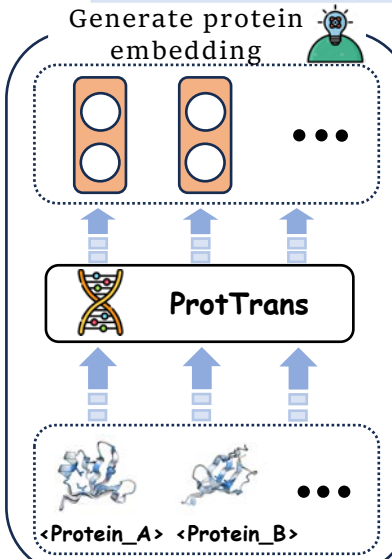
<Protein_A> acts as receptor protein, activate **<Protein_B>** . **<Protein_B>** acts as signaling protein, inhibit **<Protein_C>** . **<Protein_C>** acts as effector protein, **catalysis <Protein_D>**.
let's think it step by step. What is the relationship between **<Protein_A>** and **<Protein_D>**?

Answer:

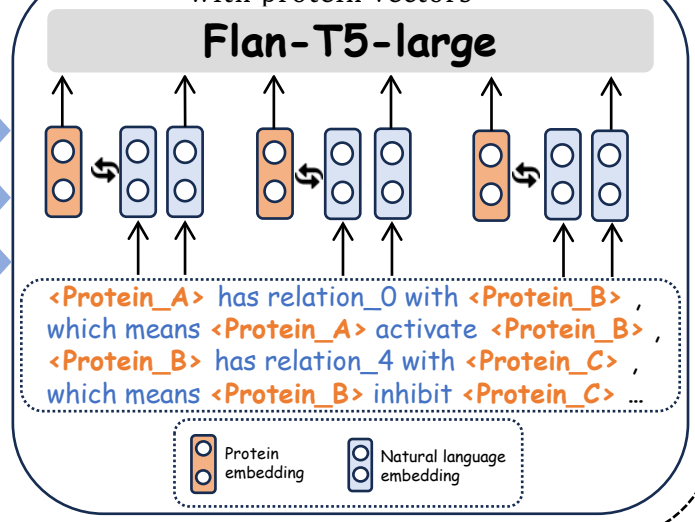
The relationship is activation.

Sector 2:Replace with ProtTrans Embedding

Generate protein embedding



Replace the T5 embedding with protein vectors



Sector 3: Instruction fine-tuning on ProLLM

Key information of Mol Dataset

Instruction: Based on the protein sequence below, estimate its function and the biological process ...

The protein sequence:
AIESVLQERDJFOWM ...

Annots: ENSP282908 (protein id) | acetate-CoA ligase activity, AMP binding... (about the protein function and biological processes of proteins)

Output: Upon inspection of the protein sequence provided, it is likely that ...

Translate into QA dataset

Question: Follow the instruction, answer my question, here is my instruction **<Instruction>**,the protein sequence is **<The protein sequence>**, the name of the protein is **<Annots>**, the protein function and biological processes of proteins is **<Annots>**.

Answer: The answer is **<Output>**.

LLM has learned relevant knowledge of biological proteins

Sector 4: Train on ProCoT format dataset

Input question

<Protein_A> acts as receptor protein, activate **<Protein_B>** . **<Protein_B>** acts as signaling protein, inhibit **<Protein_C>** . **<Protein_C>** acts as effector protein, **catalysis <Protein_D>**.
let's think it step by step. What is the relationship between **<Protein_A>** and **<Protein_D>**?
.....

Enter our model

Model's output

The relation is activation.
.....

Real answer (label)

The relation is inhibition.
.....

Negative Log Likelihood Loss